MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2018/2019

TDS3551 – DATA MANAGEMENT

(All Sections / Groups)

13 OCTOBER 2018
9:00 a.m. – 11:00 a.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 7 pages (including cover page) with 5 Questions only.

2. Attempt **ALL** questions. The distribution of the marks for each question is given. This paper carries **100 marks**.

3. Please print all your answers in the Answer Booklet provided in this examination.

Question 1 (20 Marks)

a) Sharding of a database is the partitioning of a large database into smaller separate databases that will perform faster. This is an important technique in obtaining scalability for a Relational Database Management System (RDBMS).

i. It is also known as horizontal partitioning of the database. Using an example, explain what is meant by horizontal partitioning?

[2 marks]

ii. The index/key must be spread uniformly across the shards. Explain why this is needed, paying special attention to the term "uniformly".

[2 marks]

b) In a Relational Database Management System (RDBMS), a batch queue can be used to manage high velocity of voluminous writes to the database. Without it, the writes would overwhelm the RDBMS. Explain, with the help of a diagram, how this technique works to alleviate the issue.

[4 marks]

c) "The Global Database of Events, Language and Tone (GDELT) Project monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world." (from the GDELT website). The use of open data is obvious and, in the visualization, what do the two indicators represent? (represented by the red and purple dots on the website).

[2 marks]

d) The World Bank states that "data, much like oil, has intrinsic value in that they need to be 'refined' to realise their value. When government data are made accessible and re-usable, they enable individuals, organizations and even governments themselves to innovate and collaborate". State and explain the 4 (FOUR) benefits of the adopting the Worldwide Open Data Initiative.

[8 marks]

Continued ...

e) What is the website URL of the Malaysian Open Data Initiative and in how many languages is it available in?

[2 marks]

Question 2 (20 Marks)

a) State the functionality of the following Big Data tools:

i. Sqoop
ii. Pig
iii. Flume
iv. Kafka

[4 marks]

b) State and briefly explain all 8 (EIGHT) desired properties of a Big Data Architecture.

[12 marks]

c) State and explain at least 2 (TWO) of the disadvantages of a fully incremental system as oppose to the re-computational system of the Big Data architecture?

[4 marks]

Question 3 (20 Marks)

a) Columnar Based Databases.

i. Differentiate a Columnar Based Database from a regular RDBMS and state an advantage of using a Columnar Based Database.

[2 marks]

ii. State 2 (TWO) Columnar Based Databases products.

[2 marks]

b) Illustrate, using a diagram, the inter-relationship between the Batch Layer, Speed Layer and the Serving Layer. Explain the flow of the incoming and historical data to the queries made by applications.

[6 marks]

Continued ...

c) Using the database table below, develop a fact-based model representation for the following data. State any assumptions and rationale behind your model.

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Avenue	Gotham	54
Wonder	Woman	988 Truth Way	Paradise	39
Donald	Duck	556 Quack Way	Mallard	65
Bugs	Bunny	89 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Street	Canyon	61

[6 marks]

d) For the GraphML code below, draw the graph.

```
<graphml>
 <graph id="G" edgedefault="undirected">
  <node id="1"/>
  <node id="2"/>
  <node id="3"/>
  <node id="4"/>
  <edge id="e1" source="1" target="2"/>
  <edge id="e2" source="2" target="3"/>
  <edge id="e3" directed="true" source="2" target="4" / >
  <edge id="e4" directed="false" source="3" target="1"/>
 </graph>
</graphml>
```

[4 marks]

Question 4 (20 Marks)

a) The Master Dataset is generally too large to exist on a single server and is continuously growing at a rapid pace. A file system seems to be an appropriate way to store this data that should have immutable property. Explain what is meant by immutable in this context.

[3 marks]

Continued ...

b) In a Hadoop Distributed File System (HDFS):

i. There are two categories of nodes; the DataNode and the NameNode. Give a brief explanation for each category.

[2 marks]

ii. What is the default data block size in HDFS?

[1 mark]

iii. Each data block is replicated 3 (THREE) times by default. Assuming that there are 4 data blocks represented as {A, B, C, D} and 6 DataNodes, use a diagram to show an example of how the blocks are distributed across the DataNodes (using the default number of replication).

[3 marks]

c) The MapReduce function can be explained using the following steps:

1. Read lots of data
2. **Map**
3. Partitioning, shuffle, sort, combining, merging
4. **Reduce**
5. Write the results

Give a brief explanation for the "Map" and "Reduce" stages.

[2 marks]

d) Explain why normalized data has lower performance (computational speed) than using de-normalized data.

[2 marks]

Continued ...

e) Batch view is striking a balance between the size of the precomputed views and the amount of on-the-fly computation required at query time. Assuming that data is available for daily sales of a retail store. In precomputing the views, it is decided that in addition to daily sales views, monthly and quarterly views will also be precomputed.

i. How many buckets of data is required for the 3 (THREE) granularities (daily, weekly and quarterly)?

[2 marks]

ii. What is the percentage of additional space (effort) required for create the buckets, as oppose to just storing the daily buckets?

[1 mark]

iii. For a query for sales from March 20th to July 5th, how many buckets must be accessed and what is the percentage savings as oppose to accessing it with only the daily buckets stored?

[4 marks]

Question 5 (20 Marks)

a) What is the purpose of the Speed Layer and which of the desired properties of a Big Data Architecture is specifically addressed by the Speed Layer?

[3 marks]

b) The Speed Layer stores data for a few minutes or hours. The Speed Layer produces realtime views that represent recent data and is a continuously updated at a rapid pace. Assuming that the overall system receives about 1 TeraByte of data a day where data arrives in an average size of 1 MegaByte. The Batch Layer is scheduled to generate Batch Views 8 times a day. How many times does the Speed Layer needs to do the computation (assuming re-computation method) and on average, what is the size of each of the computation?

[6 marks]

c) What does CAP (in CAP Theorem) stands for and briefly explain the CAP Theorem.

[2 marks]

Continued ...

d) Expiring the data from the Speed Layer is dependent on the Batch Layer computation and the availability of the batch views on the Serving Layer. Explain why two sets of realtime views are used to ensure that the expiring of the data from the Speed Layer will not cause issues.

[5 marks]

e) In stream processing, differentiate between One-At-A-Time and Micro-Batching processing.

[2 marks]

f) Kafka is a messaging system. Explain what is meant by this?

[2 marks]

End of Page